ANALOGIE E DIFFERENZE TRA GLI EDIFICI U1 E U6 DELL'UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA SULLA BASE DEI CONSUMI DI POTENZA ATTIVA

Davide Mancino, d.mancino1@campus.unimib.it, Matricola: 847700

Sinossi. Questo progetto ha come obiettivo quello di risalire alle principali analogie e differenze di attività, utilizzo o struttura, tra gli edifici U1 e U6 dell'Università degli studi di Milano-Bicocca a partire dalle relative serie storiche dei consumi di potenza attiva. Si è cercato di risalire alle attività didattiche e studentesche dei due edifici analizzandone i dati dei consumi di potenza attiva nei diversi mesi dell'anno, nelle differenti fasce orarie (diurne e notturne) e durante i weekend e festività. Inoltre, è stato sviluppato un sistema predittivo per i rispettivi consumi di potenza attiva, nel quale sono stati testati e validati modelli SARIMA e UCM. È stato interessante notare come, alla fine del progetto, le informazioni estratte a partire dalle serie storiche sui consumi di potenza attiva si siano rivelate verosimili ad attività, luoghi e abitudini dei rispettivi edifici che difficilmente ci si poteva aspettare di estrarre da dati su consumi di potenza attiva. Infine, per entrambi gli obiettivi si è osservato come il lockdown dovuto alla pandemia da Covid19 abbia influenzato l'andamento dei consumi in entrambi gli edifici.

Parole chiave. Time series; SARIMA; UCM; Consumi; EdificiU1/U6.

1. Introduzione. I dati analizzati in questo progetto sono stati forniti in singoli file contenenti le occorrenze dei singoli mesi e rappresentano i valori dei consumi di potenza attiva misurati nei diversi edifici ogni 15 minuti dal 2018-01-01 00:00 fino al 2020-12-31 23:45.

I formati dei file non erano organizzati in maniera omogenea e in alcuni casi erano presenti dati aggregati e formule. Il primo passo è stato quello di standardizzare i dati presenti nei vari fogli di calcolo eliminando eventuali formati e formule presenti. Successivamente, si è passati alla gestione dei dati anomali, duplicati o assenti causati principalmente dal passaggio dall'ora solare a quella legale e viceversa. In seguito alle attività di preprocessing e data cleaning si è passati ad alcune visualizzazioni relative alle serie storiche per iniziare ad analizzarne e comprenderne maggiormente l'andamento. Visualizzando i primi andamenti annuali delle serie, ci si è focalizzati nell'estrapolare informazioni interessanti rispetto alle somiglianze e differenza tra gli edifici U1 e U6. Si è proseguito con l'analisi di diversi andamenti:

- nelle diverse fasce orarie in relazione al giorno della settimana;
- settimanali;
- nelle diverse fasce orarie al variare del mese.

In modo da estrarre altre informazioni per spiegare i diversi valori tra i due edifici.

Prima di passare allo sviluppo e alla valutazione dei modelli per il sistema predittivo delle serie storiche, si è approfondito l'andamento dei dati durante il lockdown partito a marzo 2019. Successivamente, si è fatta una definizione iniziale dei modelli.

La prima tipologia di modelli sviluppata è stata quella dei modelli *SARIMA*, che prevede come precondizione che le due serie storiche siano stazionarie. Per confermare questa ipotesi sono stati utilizzati due differenti test: il *Kwiatkowski-Phillips-Schmidt-Shin* (KPSS) test e il test *Augmented Dickey-Fuller* (ADF).

Visto che inizialmente l'ipotesi di stazionarietà dei due test veniva rifiutata, si è proceduto a effettuare delle differenze di stagionalità e ottenere una serie stazionaria. Dopo aver verificato l'ipotesi di stazionarietà si è passati alla visualizzazione dei correlogrammi *ACF* (auto-correlation function) e *PACF* (partial auto-correlation function). In seguito, si è suddiviso il dataset in *train* e *validation* cercando di inserire più mesi possibili influenzati dal lockdown da pandemia Covid19. Tramite approccio *grid search* si è cercato di capire quale configurazione di parametri *p* e *q* fornisca dei *Mean Absolute Error* minori sulle previsioni dei dataset di train e validation. I modelli SARIMA sono stati costruiti sia con i dati originali che con quelli *log trasformati* e alla fine è stato scelto quello più performante

secondo il parametro di validazione del *MAE*. Per ovviare alla presenza di *stagionalità multipla*, è stato necessario aggiungere dei regressori esterni al modello SARIMA. Si è scelto, quindi, di inserire delle variabili stagionali costruite tramite serie di Fourier con periodo 168 (settimanale) e 8.760 (annuale). Individuato il modello SARIMAX più performante, sono stati analizzati i residui del modello per la validazione finale dello stesso. Oltre ai modelli SARIMA sono stati adoperati anche quelli *UCM* (Unobservable Component Model). Con approccio grid search e parametro di confronto MAE è stato scelto il *level-trend* migliore tra:

- *random walk;*
- *dconstant;*
- *local level;*
- *local linear trend;*
- *local linear deterministic trend;*
- *random walk con drift;*
- *ntrend;*
- *strend;*
- *rtrend.*

Dopo aver definito e validato i modelli SARIMA e UCM viene scelto tra i due approcci il più performante per il sistema predittivo.
Completata questa task, si è passati all'estrazione di informazioni sui due edifici facendosi guidare dai dati e traendo le dovute conclusioni.

2. Obiettivo/problema affrontato. Gli obiettivi di questo progetto sono:

1. confrontare e analizzare i consumi di potenza attiva degli edifici U1 e U6 dell'Università degli Studi di Milano-Bicocca;
2. costruire un sistema che preveda i consumi di potenza attiva;
3. estrarre informazioni dalle serie storiche per comprendere somiglianze e differenze tra i due edifici.

Questo progetto può essere molto utile per la gestione degli edifici stessi, per efficientare le risorse energetiche e per prevedere e quantificare le persone che in vari periodi dell'anno frequentano gli edifici.

3. Aspetti metodologici. I passaggi dell'approccio metodologico utilizzato sono stati:

1. studio dei dati a disposizione;
2. preprocessing e data cleaning: gestione dati missing, anomalie e sui valori uguali a 0;
3. visualizzazioni delle serie storiche per avere un quadro completo dell'andamento delle stesse;
4. visualizzazioni più specifiche riguardanti l'andamento dei dati in relazione ad archi temporali differenti, ad esempio andamento settimanale e andamento giornaliero;
5. sviluppo dei modelli;
6. ottimizzazione dei parametri dei modelli;
7. validazione dei modelli;
8. visualizzazione delle previsioni sui dati del validation set dei modelli più performanti;
9. commento sui dati ottenuti;
10. estrazione di informazioni rilevanti presenti nella serie storica.

Inoltre, è da sottolineare che l'approccio utilizzato nel gestire eventuali anomalie, all'interno dei dati originali, è stato quello di intervenire bonificando gli stessi e cercando di preservare l'andamento presente prima e dopo l'anomalia nella serie storica. Dove non è stato possibile verificare la presenza di una effettiva anomalia di misurazione, si è cercato di tenere traccia di queste evidenze senza alterarle poiché ritenute importanti ai fini degli studi sui consumi dei due edifici.

4. I dati. I dati a disposizione provengono dalle misurazioni presenti nei contatori installati negli edifici U1 e U6 dell'Università degli Studi di Milano-Bicocca effettuate ogni 15 minuti a partire dal 2018-01-01 00:00 fino al 2020-12-31 23:45. Le colonne a disposizione sono:

- POD: identificativo del contatore di energia (tipologia dato: stringa);
- DATA: giorno di rilevazione (tipologia dato: numero intero);
- ORA: valore progressivo che indica l'orario di rilevazione. Le rilevazioni sono ogni 15 minuti a partire dalle 00:00 del primo giorno del mese (tipologia dato: numero intero);
- FL_ORA_LEGALE: flag che è uguale a 1 per indicare che l'informazione dell'ora è quella solare ed è uguale a 2 se viceversa l'ora di riferimento è quella legale (tipologia dato: numero intero);

- CONSUMO_ATTIVA_PRELEVATA: valore dei kw nei 15 minuti considerati (potenza). Questa colonna rappresenta la nostra variabile di interesse (tipologia dato: numero decimale);
- CONSUMO_REATTIVA_INDUTTIVA_PRELEVATA: valore dei kw nei 15 minuti considerati (potenza) (tipologia dato: numero decimale).

Entrambi i dataset completi degli edifici U1 e U6 presentano 105.308 osservazioni.
Le rilevazioni ogni 15 minuti ci permettono di avere un dato molto granulare sui consumi di potenza attiva. Questo rappresenta un punto di forza, poiché, fornisce la possibilità di analizzare e prendere in considerazione diversi archi temporali, come le informazioni aggregate per ora, giorno, settimana, mese e anno. Allo stesso tempo rappresenta un punto di debolezza in quanto, avere delle misure con delle piccole differenze ogni 15 minuti, rende le serie storiche ricche di piccole oscillazioni intorno allo stesso valore.

Analisi/Processo di trattamento dei dati. I file forniti in origine sono stati 36 per ogni edificio. Ogni file rappresenta le occorrenze presenti in ogni singolo mese del range temporale che va dal 2018-01-01 00:00 fino al 2020-12-31 23:45.
La maggior parte di essi è presente in formato .xlsx e alcuni in formato .csv e .xltx. La prima attività di preprocessing è stata, quella di standardizzare i formati dei dati modificando i file .csv e .xltx in formato Excel (.xlsx) e di eliminare eventuali informazioni distoniche come colonne superflue o celle che presentavano formule o aggregazioni. Dopo aver effettuato queste operazioni di preprocessing, si è passati ad aggregare in un unico dataset i 36 file per edificio. L'import dei file e l'intero progetto è stato svolto in Python. Durante questa prima fase, la libreria principalmente utilizzata è stata Pandas. Successivamente, le colonne di DATA e ORA sono state trasformate in un'unica colonna denominata *DATA_ORA* la quale contiene le informazioni di data e ora delle rilevazioni in formato datetime YYYY-MM-DD HH:MM:SS.
es. DATA= *20180101*, ORA= *0* → DATA_ORA= *2018-01-01 00:00:00*

In seguito, si è passati ad analizzare eventuali problemi all'interno dei dati ad esempio:

- rilevazioni duplicate per DATA_ORA;
- rilevazioni mancanti per alcune date;
- valori rilevati uguali a 0.

Per risolvere la prima anomalia, si è proceduto identificando le rilevazioni dei consumi di potenza attiva duplicati per medesima data e ora e si è potuto osservare come tutte queste osservazioni si siano venute a creare durante il passaggio da ora legale a ora solare nei mesi di ottobre. La bonifica di questi dati è stata effettuata analizzando le singole osservazioni e recuperando solamente la rilevazione non distonica della serie storica. Le osservazioni escluse sono state eliminate.
Per gestire il secondo problema, è stato recuperato l'orario e la date di tutte le rilevazioni non presenti all'interno dei dati importati. Anche questa volta le seguenti osservazioni distoniche sono dovute al cambio di orario, stavolta presente nei mesi di marzo con il passaggio da ora solare a ora legale. In questo caso, visto che per entrambi gli edifici le rilevazioni da bonificare erano solamente 3 ('2018-03-25 02:00:00', '2019-03-31 02:00:00', '2020-03-29 02:00:00'), si è deciso di inserire come valore mancante quello rilevato durante l'ora precedente. Essendo rilevazioni notturne e non essendoci grosse differenze in consumi, è stata scelta la precedente soluzione in modo da non alterare l'andamento delle serie storiche. Prima di affrontare l'ultimo tema, i dati presenti nelle serie storiche sono stati aggregati per *ora*.
L'aggregazione delle osservazioni rilevate in una singola ora è stata fatta considerando il valore medio delle osservazioni, poiché, l'unità di misura delle stesse sono i kw.
Il terzo ed ultimo problema è stato quello di risolvere le rilevazioni di consumo di potenza attiva uguali a 0. Per questo tema si è agito in due modi differenti:

1. La prima modalità ha previsto la gestione di 2 rilevazioni isolate dove si è deciso di sostituire il valore uguale a 0 con quello rilevato l'ora precedente, senza ricondurre anche i valori molto vicino allo 0 successivi alla precedente rilevazione. In questo modo, si è comunque riusciti a tener traccia della rilevazione anomala (all'interno della Figura 1. e Figura 2. viene riportato un esempio che descrive quanto precedentemente scritto).

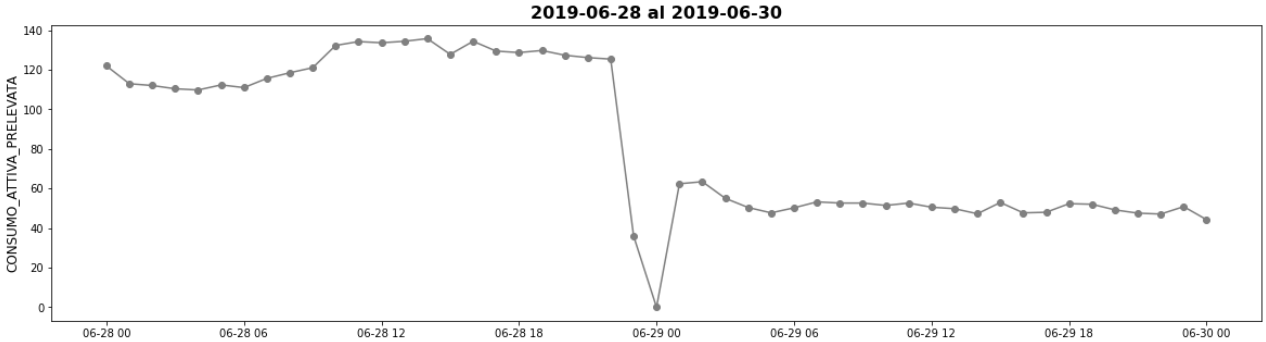

Fig. 1.: Serie storica edificio U1, il 2019-06-29 alle 00:00 presente rilevazione a 0.

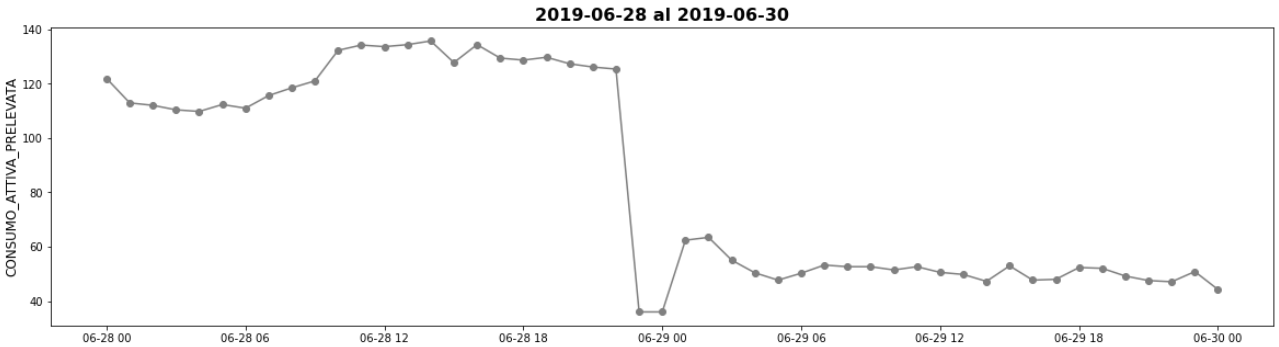

Fig. 2.: Serie storica edificio U1, dopo gestione rilevazione originaria pari a 0.

Si è scelto di sostituire la rilevazione uguale a 0 con il valore rilevato durante l'ora precedente, in modo da preservare l'andamento decrescente repentino e eliminare l'osservazione uguale a 0, la quale non avrebbe permesso una possibile trasformazione logaritmica.

2. Per entrambi gli edifici è presente una intera giornata con rilevazioni che assumono valori uguali a 0 e corrisponde alla data del 2020-07-31. In questo caso, essendo una intera giornata, si è deciso di inserire i valori medi corrispondenti alle ore di rilevazione tra il giorno precedente e quello successivo. È stata scelta questa soluzione perché come si vedrà successivamente è presente una forte stagionalità mensile, settimanale e giornaliera. Si è pensato quindi che questa soluzione potesse essere la migliore per preservare l'andamento della serie storica. All'interno delle Figure 3. e 4. viene visualizzata la serie storica con i dati originari e quelli successivi alla bonifica dei valori uguali a 0.

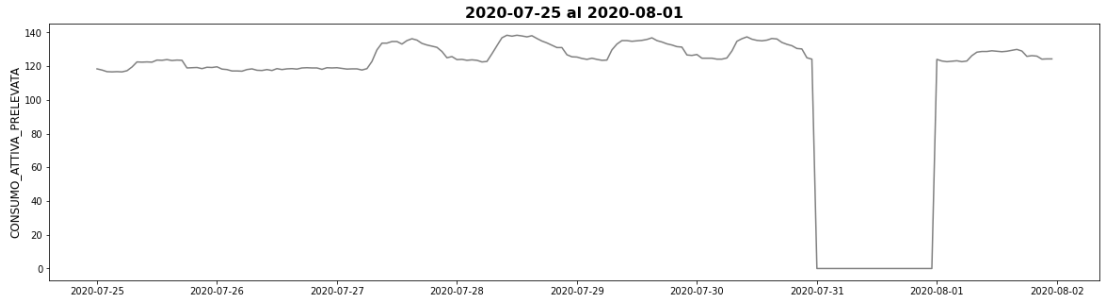

Fig. 3.: Serie storica edificio U6, i dati originari della giornata del 2020-07-31.

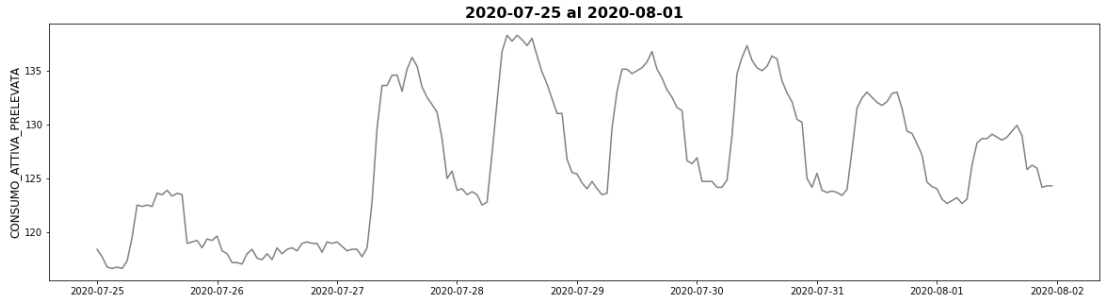

Fig. 4.: Serie storica edificio U6, i dati post bonifica della giornata del 2020-07-31.

Completate queste operazioni di data cleaning e preprocessing, sono stati salvati due file .csv contenenti il dataset completo con le osservazioni degli edifici U1 e U6.

Fatto ciò, ci si è focalizzati sulle visualizzazioni iniziali per estrarre possibili archi temporali critici, per confrontare gli andamenti tra i due edifici e per provare a giustificarne somiglianze ed eventuali differenze.

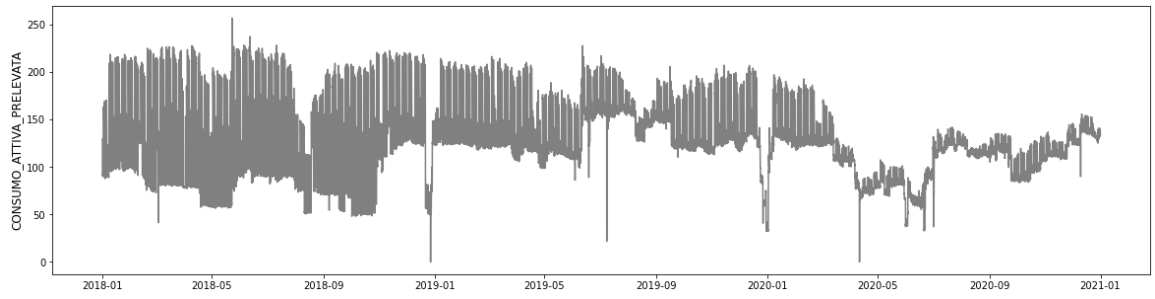

Fig. 5.: Serie storica completa edificio U6.

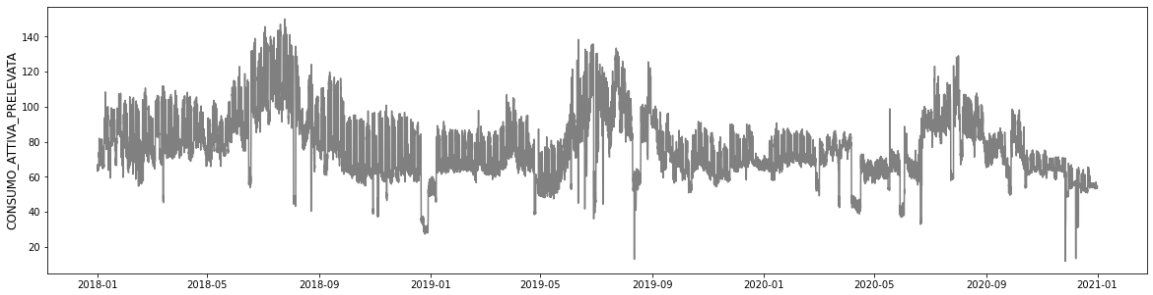

Fig. 6.: Serie storica completa edificio U1.

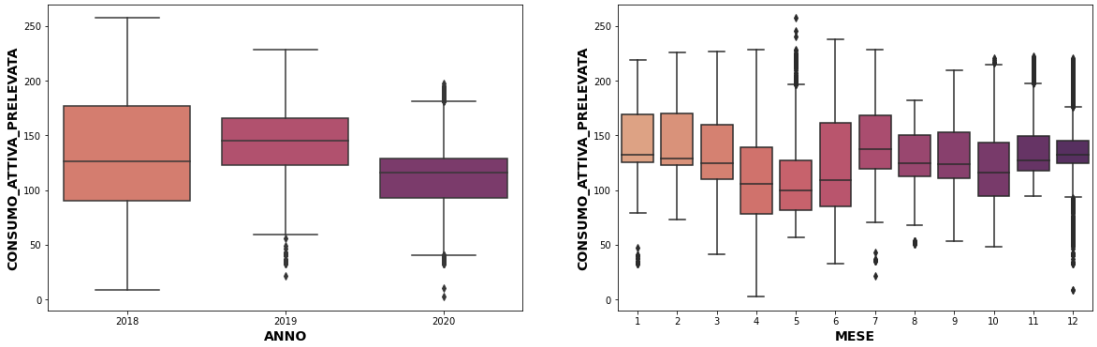

Fig. 7.: Serie storica edificio U6, Box plot annuali e mensili.

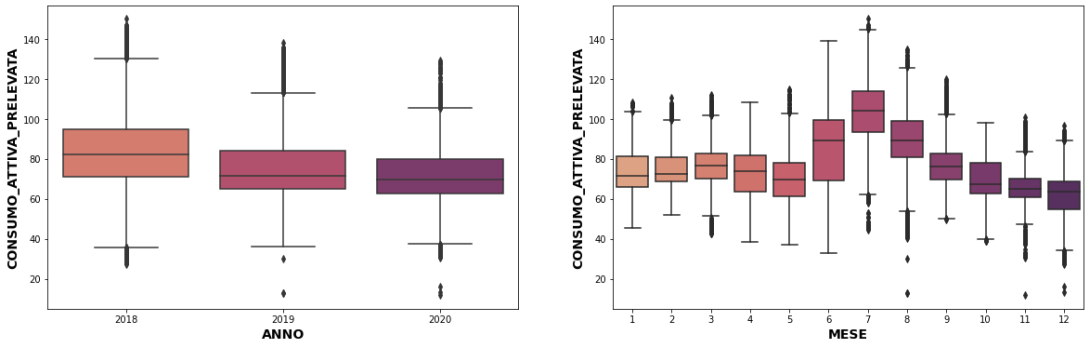

Fig. 8.: Serie storica edificio U1, Box plot annuali e mensili.

Le figure 5. e 6. rappresentano le serie storiche complete dei consumi di potenza attive degli edifici U1 e U6 a partire dal 2018-01-01 00:00 fino al 2020-12-31 23:45. Dalle immagini emerge come siano presenti alcuni picchi repentini decrescenti rappresentati dai valori outlier presenti nei box plot delle figure 8. e 9.
Alcune date in cui si manifestano questi picchi si possono ricondurre a giornate di festività nazionale. Inoltre, i box plot delle figure 8. e 9. ci forniscono delle informazioni sull'andamento annuale e mensile delle serie storiche.
Per quanto riguarda l'edificio U1 possiamo osservare un trend decrescente dei consumi di potenza attiva tra il 2018 e 2020 (ricordiamo anche che durante il 2020 è stato presente un lockdown dovuto a pandemia da Covid19), mentre i box plot mensili di entrambi gli edifici ci confermano come il mese di luglio sia quello in cui si registrano più consumi di potenza attiva. Questo è da attribuire probabilmente, all'attivazione degli impianti di condizionamento.
I box plot mensili ci danno già una indicazione di un possibile andamento annuale confermato dai grafici presenti nella figura 9. Questo andamento è più visibile confrontando i box plot e la stagionalità annuale dell'edificio U1 piuttosto che confrontando quelli dell'edificio U6. Inoltre, già a partire dal grafico dei box plot, si può osservare come i valori dei consumi di potenza attiva massima dell'edificio U1 siano inferiori rispetto a quelli dell'edificio U6.
Oltre all'analisi sulla stagionalità annuale sono state fatte delle analisi su:

1. andamento dei giorni della settimana (Figura 10.);
2. andamento dei consumi per specifica ora al variare dei giorni della settimana (Figura 11.);
3. andamento dei consumi per specifica ora al variare dei mesi (Figura 12.).

Da questi grafici si possono estrarre numerose informazioni riguardo i due differenti edifici. Ad esempio, nell'edificio U6 si trovano dei picchi di consumi di potenza attiva a luglio, gennaio, febbraio e novembre, mentre nell'edificio U1 abbiamo una crescita repentina solo per il mese di luglio (Figura 9.). Inoltre, osservando gli andamenti presenti all'interno della Figura 10. (che rappresentano l'andamento settimanale), si può verificare come i consumi medi giornalieri di

potenza attiva siano molto simili dal lunedì al venerdì, mentre rilevano una repentina decrescita durante i weekend. Questo è giustificato dal fatto che le attività universitarie sono concentrate maggiormente dal lunedì al venerdì, diversamente dai weekend dove in U6 restano operativi solamente uffici ed aule studio per qualche ora della giornate, mentre l'U1 resta chiuso. Un'altra conferma a favore di questa tesi si può ritrovare all'interno della figura 11, nella quale si osserva l'andamento dei consumi orari per giorno della settimana. La stessa, infatti, oltre a dimostrare come i valori di consumo registrati tra sabato e domenica siano inferiori rispetto a quelli presenti nel resto della settimana, evidenzia come nelle ore notturne (dalle 22:00 alle 03:00) vengano registrati i valori più bassi. Inoltre, dalle 05:00 alle 13:00 si può osservare una forte crescita dei consumi da potenza attiva. Per l'edificio U6 dal lunedì al venerdì il picco massimo si raggiunge alle 13:00, mentre nell'edificio U1 il picco si registra alle 15:00. L'alto consumo di potenza attiva registrato in U6 alle 13:00 potrebbe essere attribuito all'attivazione del servizio mensa, mentre in U1 potrebbe essere attribuito all'inizio delle lezioni. Dall'analisi si può evincere come in U1 i valori inizino a decrescere prima rispetto all'U6, probabilmente per l'orario di chiusura (anticipato alle ore 20:00) e per le modalità di utilizzo dell'edificio (finalizzato alle lezioni) differentemente dall'edificio U6 utilizzato anche per studio singolo o di gruppo grazie alla presenza di aule studio, biblioteca, laboratori e aula magna. Inoltre, l'orario di chiusura dell'edificio U6 è alle 22:00. È anche da sottolineare che il sabato, l'edificio U6 registra un andamento crescente rispetto ai dati registrati per l'edificio U1, poiché resta aperto agli studenti (fino alle ore 14:00) e viene utilizzato per alcuni laboratori.

La figura 12 presenta gli andamenti orari registrati nei vari mesi. Il grafico indica come i mesi più caldi, rappresentati da un colore tendente al porpora, siano quelli che presentano una crescita repentina con picchi massimi maggiori verso l'ora di punta. Questo potrebbe essere giustificato dall'azionamento dei sistemi di condizionamento.

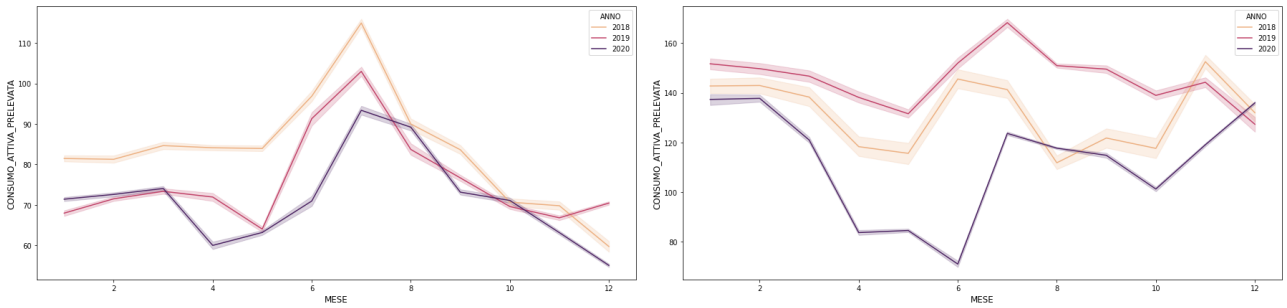

Fig. 9.: a sinistra stagionalità annuale edificio U1, a destra stagionalità annuale edificio U6

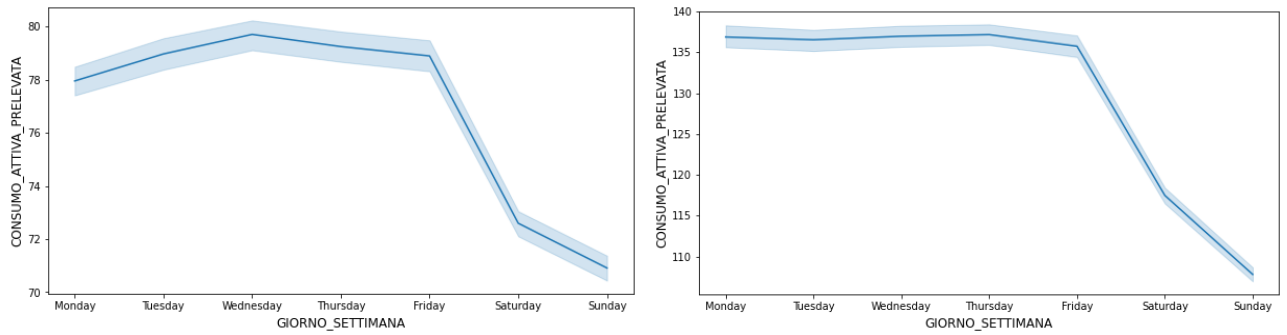

Fig. 10.: a sinistra andamento settimanale edificio U1, a destra andamento settimanale edificio U6

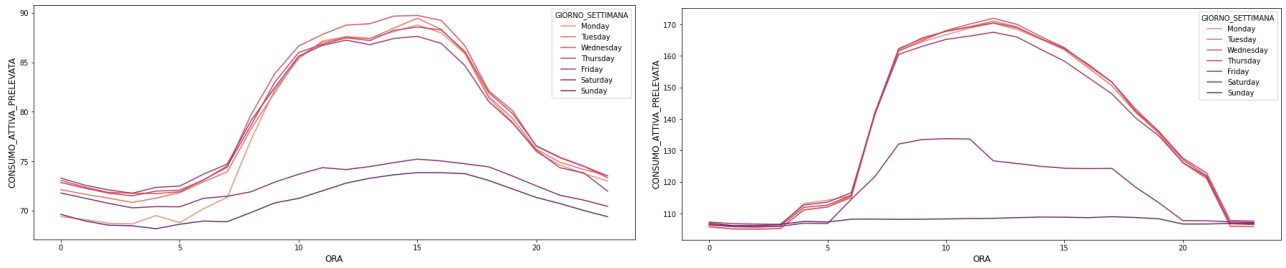

Fig. 11.: a sinistra andamento ogni ora per giorno della settimana edificio U1, a destra rispettivo edificio U6.

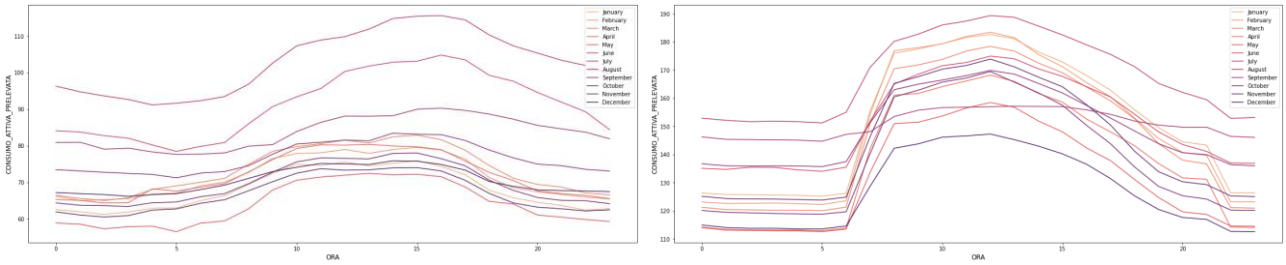

Fig. 12.: a sinistra andamento ogni ora per mese edificio U1, a destra rispettivo edificio U6.

La figura 13 rappresenta un approfondimento sull'influenza che la pandemia da Covid19 ha avuto sui dati. Il lockdown è iniziato il 2020-03-09, da questa data si può osservare una repentina decrescita sui consumi di potenza attiva. Prima di commentare la figura, è da sottolineare la presenza di dati errati per il mese di giugno 2020. Infatti, i dati relativi a questo mese per l'edificio U6 corrispondo alle rilevazioni fatte per l'edificio U1.

Si è deciso di non intervenire nella serie storica dell'edificio U6, poiché, l'influenza della pandemia da Covid19 rende difficile una possibile stima. Per rappresentare il periodo della pandemia è stata creata una variabile step "Covid19" che verrà utilizzata al momento dello sviluppo dei modelli. Osservando la figura 13. si può evincere come l'andamento dei consumi da potenza attiva sia stato molto più influenzato dalla pandemia per l'edificio U6 che per l'edificio U1.

Fig. 13: in alto focus inizio lockdown per pandemia Covid19 per l'edificio U1, in basso per l'edificio U6

Completate queste altre operazioni si può passare alle fasi di definizione, sviluppo e validazioni dei modelli per un sistema di previsione dei consumi di potenza attiva.

Modelli: Per lo sviluppo di un sistema predittivo sono stati testati modelli SARIMA e modelli UCM. La prima operazione è stata quella di definire un dataset di train e uno di validation. Il dataset originario è stato suddiviso nel seguente modo:

- *Dataset di Train*: 23.376 osservazioni, dal 2018-01-01 00:00 al 2020-08-31 23:00.
- *Dataset di Validation*: 2'928 osservazioni, dal 2020-09-01 00:00 al 2020-12-31 23:00.

È stato scelto di utilizzare gli ultimi quattro mesi delle serie storiche per permettere ai modelli di validare dati non troppo influenzati dalla pandemia da Covid19. I modelli così potranno avere a disposizioni i primi mesi del lockdown partito a marzo 2020 per apprenderne l'andamento e cercare di restituire delle previsioni consone al periodo.

SARIMA: dopo aver suddiviso il dataset originario in train e validation si è iniziato a sviluppare la prima tipologia di modelli, quelli SARIMA. Uno dei requisiti per i modelli ARIMA è che le serie temporali siano stazionarie. Per trasformare una serie non stazionaria in una stazionaria è possibile effettuare trasformazioni logaritmiche (per stabilizzarne la varianza) o differenze (per stabilizzarne la media). Prima di effettuare differenze o trasformazioni logaritmiche ci si è forniti dell'utilizzo dei test Augmented Dickey-Fuller (*ADF*) e Kwiatkowski-Phillips-Schmidt-Shin (*KPSS*) per verificare se le serie storiche siano stazionarie oppure no.
Con un livello di significatività dello 0,95 , la serie risulta essere stazionaria secondo il test ADF ma non per quello KPSS (Tabella 1.), per questo motivo si è deciso di applicare una differenza stagionale alla serie e successivamente a questa azione la serie risulta essere stazionaria (Tabella 2.).

```
Results of Dickey-Fuller Test:
Test Statistic       -9.645913e+00
p-value               1.478069e-16
Lags Used             4.700000e+01
Critical Value (1%)  -3.430630e+00
Critical Value (5%)  -2.861664e+00
Critical Value (10%) -2.566836e+00

Results of KPSS Test:
Test Statistic            5.779102
p-value                   0.010000
Lags Used                47.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```

Tabella 1: edificio U6, in alto test iniziale ADF, in basso test iniziale KPSS.

Applicando una differenza stagionale la serie storica risulterà stazionaria come è possibile evincerlo dalla tabella 2. che rappresenta i risultati presenti nei successivi test ADF e KPSS.

```
Results of Dickey-Fuller Test:
Test Statistic          -28.872752
p-value                   0.000000
Lags Used                47.000000
Critical Value (1%)      -3.430631
Critical Value (5%)      -2.861664
Critical Value (10%)     -2.566836

Results of KPSS Test:
Test Statistic            0.003894
p-value                   0.100000
Lags Used                47.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```

Tabella 2: edificio U6, in alto test iniziale ADF, in basso test iniziale KPSS dopo aver applicato una differenza stagionale.

Invece, dal correlogramma PACF successivo alla differenza stagionale (Figura 14.) possiamo intuire come un buon modello possa essere uno che abbia come parametri p=1 o p=2.

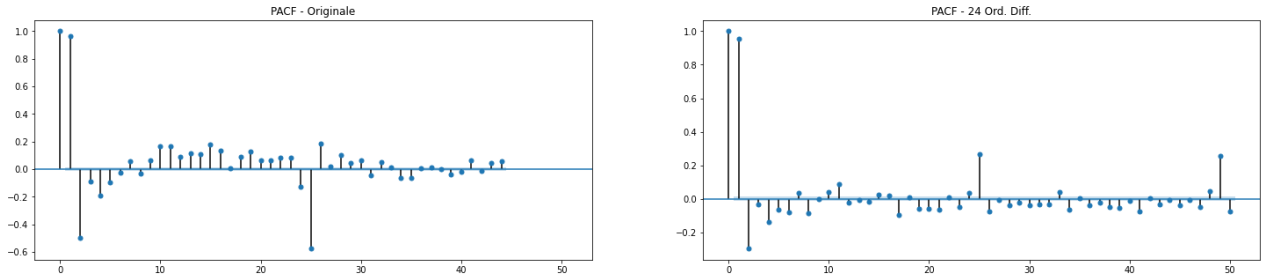

Fig. 14: edificio U6, a sinistra PACF serie storica originale, a destra PACF dopo differenza stagionale della serie storica.

Per determinare i migliori valori per i parametri non stagionali *p* e *q*, è stato utilizzato un approccio *grid search* dove sono stati testati sia i modelli con i dati originali (Tabella 3.) che quelli dopo una trasformazione logaritmica (Tabella 4.).

MODELLO	MAE Train	MAE Validation	AIC
SARIMA(0,0,0)(1,1,1)24	14,3	12,9	212.616,0
SARIMA(0,0,1)(1,1,1)24	8,4	12,9	186.178,0
SARIMA(0,0,2)(1,1,1)24	6,1	12,9	172.395,0
SARIMA(1,0,0)(1,1,1)24	3,1	12,9	151.746,0
SARIMA(1,0,1)(1,1,1)24	3,0	12,9	149.649,0
SARIMA(1,0,2)(1,1,1)24	3,0	12,9	149.597,0
SARIMA(2,0,0)(1,1,1)24	3.0	12,9	149.498,0
SARIMA(2,0,1)(1,1,1)24	3,0	12,9	149.364,0
SARIMA(2,0,2)(1,1,1)24	3,0	12,9	149.121,0

Tabella 3: edificio U6, performance modelli SARIMA con dati originali al variare dei parametri p e q.

MODELLO	MAE Train	MAE Validation	AIC
SARIMA(0,0,0)(1,1,1)24	14,8	13,0	-9.241,0
SARIMA(0,0,1)(1,1,1)24	8,7	13,0	-33.602,0
SARIMA(0,0,2)(1,1,1)24	6,5	13,0	-44.854,0
SARIMA(1,0,0)(1,1,1)24	3,1	12,9	-62.894,0
SARIMA(1,0,1)(1,1,1)24	3,0	12,9	-63.553,0
SARIMA(1,0,2)(1,1,1)24	3,1	12,9	-63.613,0
SARIMA(2,0,0)(1,1,1)24	3,0	12,9	-63.475,0
SARIMA(2,0,1)(1,1,1)24	3,1	12,9	-63.685,0
SARIMA(2,0,2)(1,1,1)24	3,1	12,9	-63.678,0

Tabella 4: edificio U6, performance modelli SARIMA con dati log-trasformati al variare dei parametri p e q.

Come è possibile evincere dalle tabelle 3. e 4. i valori di MAE non differiscono di molto tra i dati originali e quelli con trasformazione logaritmica. Viene così scelto di non applicare nessuna trasformazione logaritmica e di selezionare come miglior modello in termini di AIC e MAE il SARIMA(2, 0, 2)(1, 1, 1)24 per l'edificio U6.

Le stesse operazioni vengono effettuate per l'edificio U1 dove anche qui sono state effettuate le stesse operazioni ma stavolta scegliendo di inserire una stagionalità mensile (12) selezionando alla fine un modello SARIMA(1, 0, 1)(1, 1, 1)12.

Come è possibile osservare all'interno delle figura 15., questo modello non è molto soddisfacente, poiché durante la previsione dei dati di validazione riesce a cogliere la stagionalità giornaliera ma non quella settimanale e annuale. Per ovviare a questo problema di stagionalità multipla, si è deciso di utilizzare un modello SARIMAX.

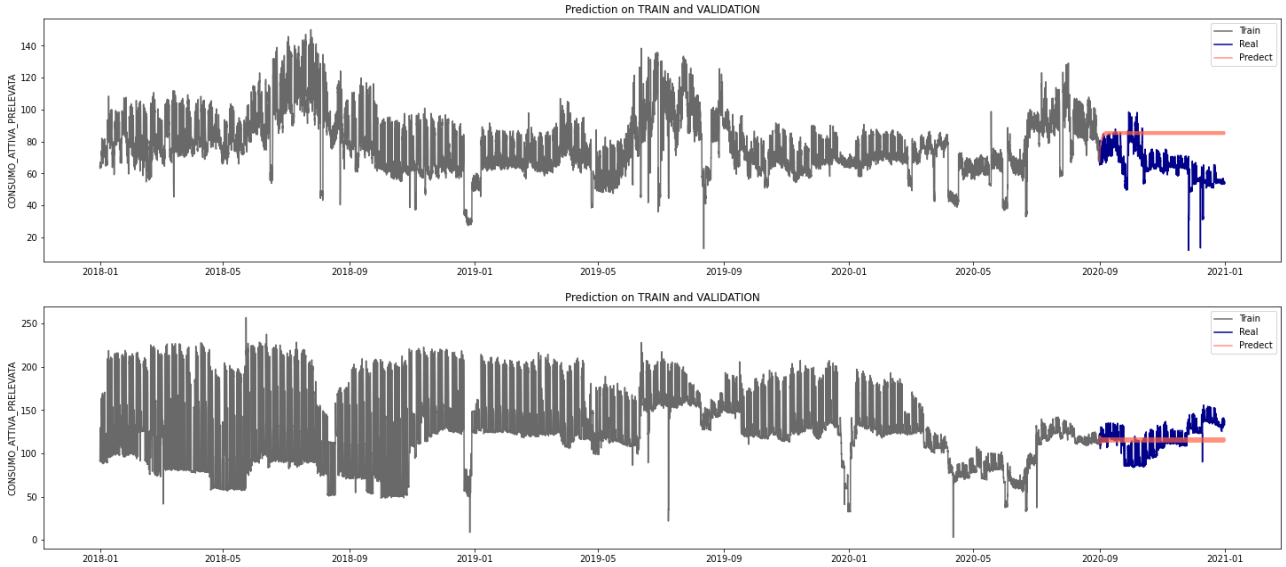

Fig. 15: in alto modello SARIMA(1, 0, 1)(1, 1, 1)12 edificio U1, in basso modello SARIMA(2, 0, 2)(1, 1, 1)24 edificio U6

Vengono inserite delle variabili stagionali costruite tramite serie di Fourier con periodo 168 (settimanale) e 8.760 (annuale). Per scegliere i parametri più performanti viene utilizzato sempre un modello grid search. Dopo questo approccio si è deciso di utilizzare il modello SARIMAX(2,0,2)(1,1,1)24 con 10 armoniche per la stagionalità settimanale e 5 per quella annuale. Inoltre, vengono effettuate delle prove aggiungendo anche la variabile step 'Covid19' precedentemente costruita senza rilevare miglioramenti. Si decide quindi di non utilizzarla. Con questo modello rileviamo un MAE sul dataset di train pari a 6,7, un MAE sul dataset di validation pari a 14,4 e un AIC pari a 2.580.742,0 . Osservando la previsione sul dataset di validation, possiamo notare come oltre la stagionalità giornaliera, adesso, il modello segue anche la stagionalità settimanale e annuale. È però da rilevare un problema di previsione riguardante i valori presenti i sabati e le domeniche. Questo problema potrebbe essere riconducibile all' influenza della pandemia da Covid19 sui dati di validation. Infatti, le differenze tra i consumi registrati durante i giorni che vanno dal lunedì al venerdì e quelli registrati durante il weekend, sembrano essere inferiori rispetto ai periodi pre-Covid19. Questo problema rende la previsione distante dai valori reali durante le giornate dei fine settimana (Figura 16. e Figura 17.).

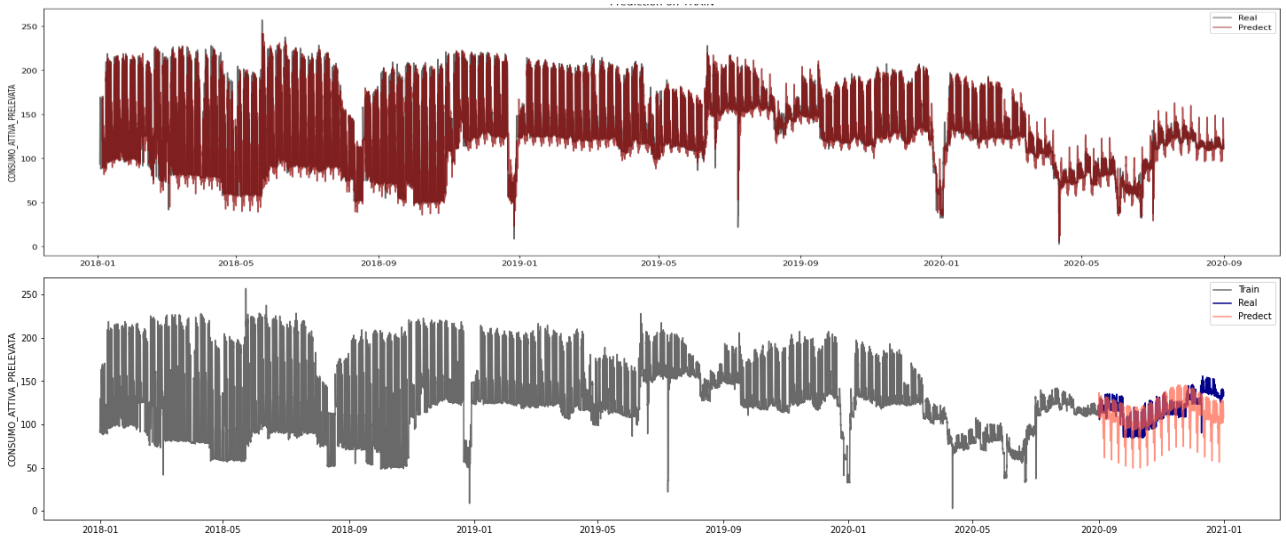

Fig. 16: in alto modello SARIMAX(2, 0, 2)(1, 1, 1)24 edificio U6 confronto dati reali e predetti sul dataset di train e in basso confronto tra i dati reali e quelli predetti sul dataset di validation.

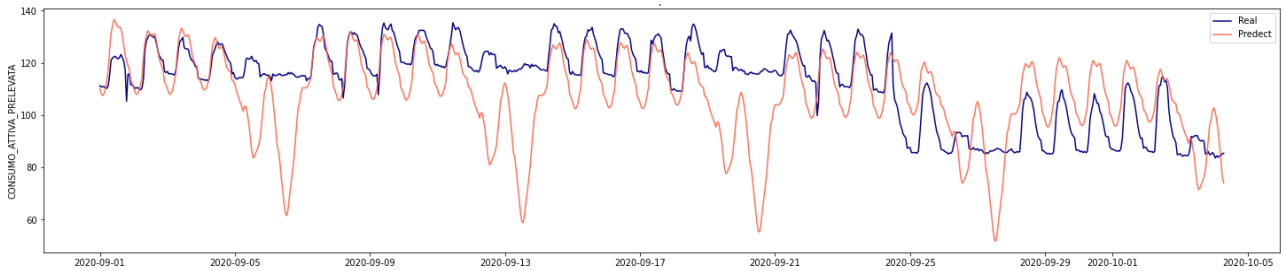

Fig. 17: modello SARIMAX(2,0,2)(1,1,1)24 edificio U6, focus sulla prima parte del dataset di validation.

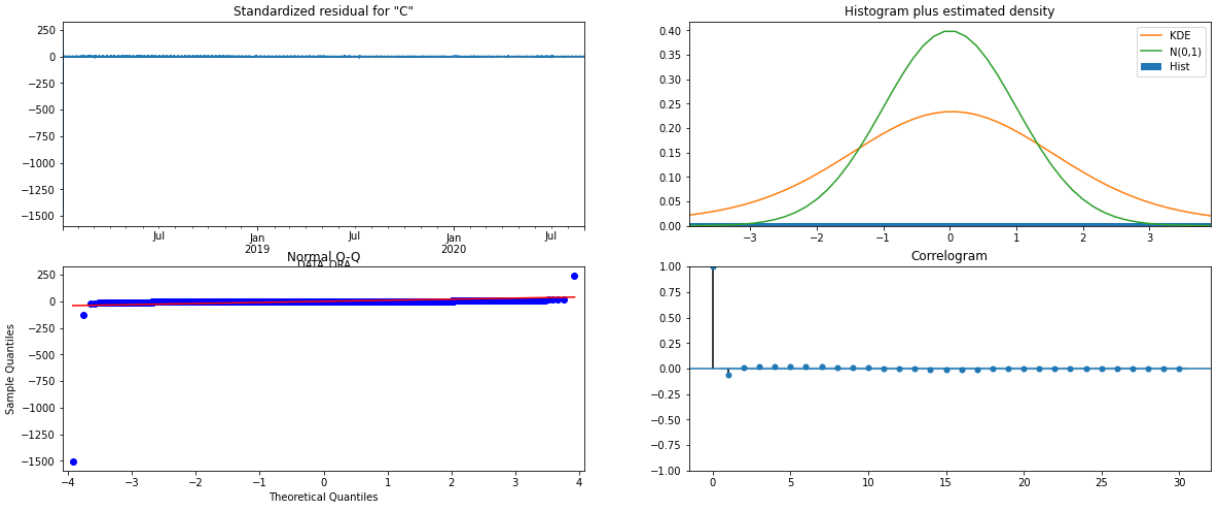

Fig. 18: edificio U6, analisi residui modello SARIMAX(2,0,2)(1,1,1)24.

Dopo aver completato queste operazioni bisognerà validare il modello andando ad osservare le caratteristiche dei residui generati. Per fare ciò sfruttiamo una funzione presente all'interno della libreria statsmodels di Python. Come è possibile osservare nella figura 18. i residui risultano essere incorrelati e la distribuzione degli stessi risulta approssimativamente normale. Nel complesso, il modello può essere validato.

UCM Oltre ai modelli SARIMA sono stati sviluppati e testati anche i modelli UCM. L'approccio metodologico è stato molto simile a quello utilizzato per validare i modelli SARIMA. È stato

suddiviso il dataset originale in train e validation, inserendo nel dataset di validation gli ultimi 4 mesi, ovvero, dall'1 settembre 2020 al 31 dicembre 2020. Sono state incluse le componenti di stagionalità e di level-trend. Per determinare il level-trend migliore è stato utilizzato un approccio grid search. Dai dati presenti nella tabella 5. riusciamo ad individuare nel random walk il miglior tipo di level-trend per valori di MAE nel dataset di train e di validation. Nella tabella 6. sono presenti i dati con la presenza della variabile step 'Covid19' dove però non sono presenti miglioramenti.
La stagionalità giornaliera è rappresentata tramite variabile dummy, mentre quella settimanale tramite 15 serie di Fourier e quella annuale da 5 serie di Fourier.
Il modello selezionato per l'edificio U6 è il random walk. Nelle figure 18. e 19. è possibile osservare le previsioni sui dati di train e di validation. Questo modello produce un MAE sul train di 3,4 e 12,0 il MAE di validation.

Per l'edificio U1, il modello UCM migliore è il random walk con drift. Dopo aver effettuato le stesse operazioni il modello ha prodotto un MAE sul train di 2,1 e un MAE sul dataset di validation di 13,1 (Figura 20.).

Level-trend	MAE Train	MAE Validation
rwalk	3,7	32,5
dconstant	5,9	47,6
llevel	24,6	109,5
lldtrend	3,7	32,5
rwdrift	3,7	40,4
lltrend	3,7	40,4
strend	4,6	1.013,5
rtrend	4,6	1.809,2

Tabella 5: edificio U6, performance level-trend modelli UCM con dati originali senza variabile covid19.

level-trend	MAE Train	MAE Validation
rwalk	3,7	32,5
llevel	3,7	32,5
lldtrend	3,7	40,4
rwdrift	3,7	40,4

Tabella 6: edificio U6, performance level-trend modelli UCM con variabile covid19.

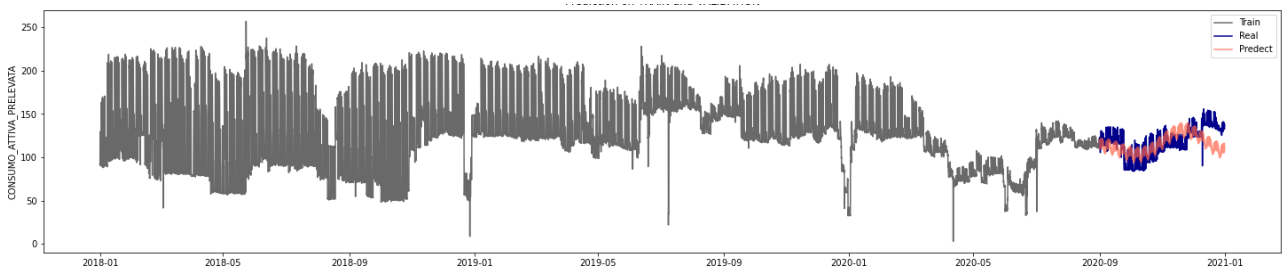

Fig. 18: modello UCM random walk con 15 serie di Fourier per stagionalità settimanale e 5 per l'annuale serie storica dell'edificio U6 confronto dati reali e predetti sul dataset di validation.

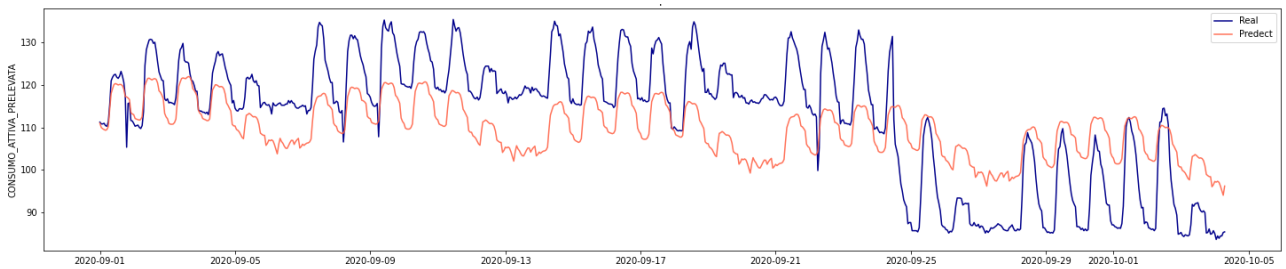

Fig. 19: edificio U6, focus su previsione dati di validation modello UCM random walk con 15 serie di Fourier per stagionalità settimanale e 5 per quella annuale.

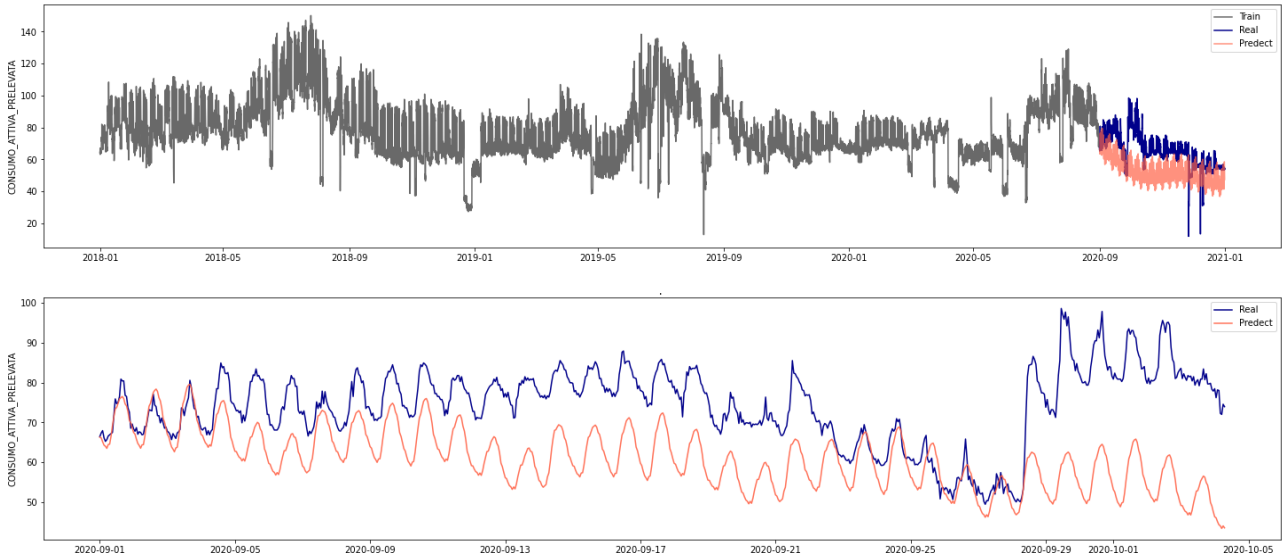

Fig. 20: edificio U1, in altro previsione sui dati di validation modello UCM random walk, in basso focus su previsione dati di validation.

Risultati. Osservando i risultati prodotti dai vari modelli, possiamo affermare che (come confermato dalla tabella 7.) per l'edificio U6 il modello più performante è stato il modello UCM random walk, mentre per l'edificio U1 il più performante è stato il modello SARIMAX(1,0,1)(1,1,1)12 (Figura 21.). È da sottolineare come sia possibile migliorare il modello SARIMAX per l'edificio U6, ma non è stato possibile effettuare troppi test perché ognuno necessitava di ingenti quantità di tempo per ritestare il modello.

Edificio	Modello	MAE Train	MAE Validation
U6	rwalk	3,4	13,9
U6	SARIMAX(2,0,2)(1,1,1)24	6,6	14,4
U1	rwdrift	2,1	13,1
U1	SARIMAX(1,0,1)(1,1,1)12	2,1	9,4

Tabella 7: Modelli più performanti per gli edifici U6 ed U1.

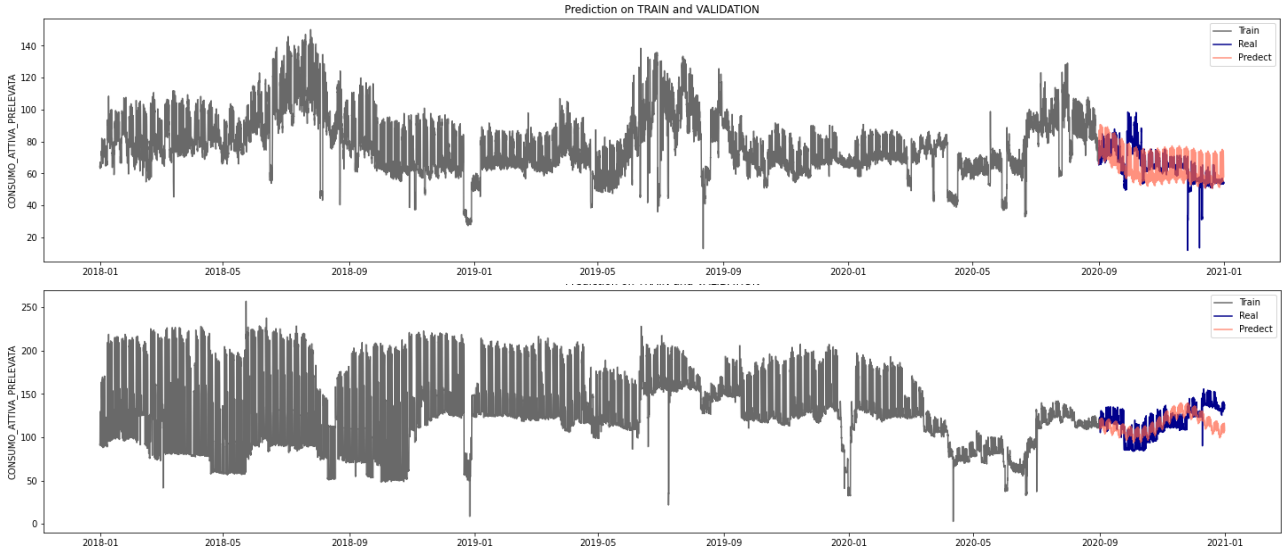

Fig. 21: in alto modello migliore per l'edificio U1 (SARIMAX(1, 0, 1)(1, 1, 1)12), in basso modello migliore per l'edificio U6 (random walk).

Per le differenze che sono emerse tra gli edifici U1 ed U6 possiamo affermare con certezza dopo le analisi effettuate nel progetto che l'edificio U6 è il più frequentato grazie al suo orario di chiusura prolungato (chiusura ore 22:00), all'apertura del sabato fino alle 14:00 mentre l'U1 resta chiuso. Possiamo affermare come i consumi di potenza attiva mediamente siano maggiori per l'edificio U6 rispetto all'edificio U1, questo anche giustificato dal fatto che l'edificio U6 è strutturalmente più grande e presenta diversi utilizzi (lezioni, aule studio, mensa, biblioteca, bar e aula magna). L'edifico U6 ospita i dipartimenti di Giurisprudenza, Economia, Psicologia e Scienze Umane per la Formazione. L'edificio U1, situato in Piazza della Scienza ospita i dipartimenti delle facoltà scientifiche ma insieme ai limitrofi edifici U2, U3 ed U4. Inoltre, le aule adibite a didattiche e studio disponibili in U1 sono 13 per 1.300 posti mentre per l'edificio U6 sono 46 per 5.311 posti (questo conferma che il valore medio giornaliero dei consumi di potenza attiva sia superiore per l'edificio U6 rispetto a quello dell'edificio U1). Inoltre, all'interno dell'edificio U6 è presente un'aula magna da 920 posti totali, nella quale vengono effettuate conferenze e cerimonie di laurea, questo potrebbe giustificare i picchi di consumo che oltre a luglio si registrano a gennaio, febbraio e novembre. I picchi presenti in U6 delle 13:00 possono essere giustificati dalla presenza di una mensa e di due bar (uno presente al primo piano e un altro al piano -1), mentre in U1 sono solamente disponibili dei distributori automatici. Infine, si deve sottolineare come la pandemia da Covid19 abbiamo influenzato molto di più in termini di consumi l'edificio U6 rispetto all'U1. Questo possibilmente riconducibile al fatto che l'edificio U6 sia quello più frequentato dagli studenti che con il lockdown non potevano raggiungere l'Università e di conseguenza sfruttare gli spazi interni ad essa.

Conclusione e possibili sviluppi. Questo progetto ha portato a definire un possibile sistema predittivo dei dati dei consumi di potenza attiva per gli edifici U6 e U1. Inoltre, l'analisi degli andamenti degli stessi ha permesso di risalire a numerose informazioni che differenziano i due edifici e ne mettono in risalto le differenze strutturali, di capienza e di attività. Gli sviluppi futuri sono da concentrarsi su un miglioramento dei modelli SARIMAX, gestire gli outlier presenti nelle due serie storiche e riflettere su una possibile gestione dei periodi di festività con delle variabili dummy stagionali.

Riferimenti bibliografici.

Dati capienza edifici UNIMIB: https://www.unimib.it/amministrazione-trasparente/altri-contenuti/ateneo-cifre/dati-sulle-infrastrutture

Dati capienza aula magna edificio U6 UNIMIB: https://www.unimib.it/node/13759

Forecasting Time Series Data with Multiple Seasonal Periods: https://tanzu.vmware.com/content/blog/forecasting-time-series-data-with-multiple-seasonal-periods

Hourly electricity demand forecasting using Fourier analysis with feedback: https://www.sciencedirect.com/science/article/pii/S2211467X20300778

How to Grid Search SARIMA Hyperparameters for Time Series Forecasting: https://machinelearningmastery.com/how-to-grid-search-sarima-model-hyperparameters-for-time-series-forecasting-in-python/

Marco Fattore, *Fundamentals of Time Series Analysis for the Working Data Scientist – Draft* (2020)

Servizi mensa e bar UNIMIB: https://www.unimib.it/servizi/campus-bicocca/mense